图书情报工作
第 64 卷 第 11 期 2020 年 6 月



# 专利视角下融合多属性的技术创新主题挖掘方法*
## ——以芯片领域专利为例

李慧 玄洪升
西安电子科技大学经济与管理学院 西安 710126

摘 要:[目的/意义]使用融合多属性的量化方法,快速且有效地挖掘出领域内多个技术创新主题,为技术创新方向的确定提供借鉴。[方法/过程]将 LDA(Latent Dirichlet Allocation)主题模型与专利价值评价指标相结合,提出一种挖掘技术创新主题的量化方法。首先,综合运用 TF-IDF、困惑度和四分位数法构建领域专利的 LDA 主题模型。然后,利用 LDA 输出的概率分布矩阵,结合专利价值评价指标(权利要求和 IPC),构建量化指标体系。接着,选取芯片专利进行验证实验,计算量化指标并运用热力图对其可视化,识别出技术创新主题。最后,基于专利、LDA 的输出矩阵、创新主题和量化指标之间的映射关系,进行专利筛选和技术创新主题的合理标记。[结果/结论]通过邀请微电子领域专家和参考最新国内外芯片技术两种方式对实验结果进行评估,结果表明:融合多属性的领域技术创新主题挖掘方法能够快速且有效地挖掘出多个技术创新主题,在实践层面可以更好地为相关领域企业和科技工作者发现技术创新主题提供思路。

关键词:专利 困惑度 潜在狄利克雷分布 量化指标体系 技术创新主题
分类号:G250
DOI:10.13266/j.issn.0252-3116.2020.11.011



## 1 引言

随着我国加快建设创新型国家战略的提出,科技创新被摆在国家发展全局的核心位置,成为提高社会生产力和综合国力的战略支撑。在这种国家战略背景下,科技情报因其能为科技创新提供全局性、战略性的决策支撑显得尤为重要。因此,选择何种数据源作为获取前沿技术创新情报的可靠途径就成为学者们研究的对象。

在众多科技文献数据中,专利数据是学术界最常用的技术创新衡量指标[1],因此,专利数据成为获取技术创新情报的有效途径之一。J. Schmookle[2]在研究中使用专利数据表征技术创新,以 1836-1957 年美国铁路运输业、农业、造纸业和石油加工业的专利申请量考察这 4 个行业的技术革新情况。Z. Griliches[3]提出专利数据是技术创新的重要信息来源,并且在完整性和技术创新信息披露等方面,专利数据具有其他指标不可替代的优势。U. Schmoch[4]研究发现专利数据可以用于分析特定领域的技术创新水平,弥补了研发预算和研发人员等指标通常只在总体层面上进行统计的不足。经济合作与发展组织(Organization for Economic Co-operation and Development, OECD)[5]指出,专利和专利数据不仅能揭示发明所属的技术领域,还能掌握发明申请人、授让人和发明人的相关信息。赵阳等[6]指出,随着专利数据的飞速增长,学术界用于挖掘专利技术创新信息的手段(专利检索、专利地图、专利引文、专利网络、专利文本挖掘)也在快速更新迭代。

国内外学者充分认识到专利数据衡量技术创新的可行性[1]。因此,他们就如何利用专利数据高效、准确地识别出技术主题展开了大量的研究工作。根据相关研究,技术主题识别主要分为两个方向:一是基于专利引用特征,二是基于专利文本内容特征。其中,基于专利引用特征的技术主题识别方法较早受到学者的关注。C. Choi 等[7]构建了专利引文网络,利用主路径分

* 本文系中央高校基本科研业务费专项资金项目“专利视角下的技术创新主题发现与趋势预测”(项目编号:JB190610)和国家自然科学基金青年项目“基于公众网络参与的民生公共政策第三方动态评估机理与方法研究”(项目编号:71503195)研究成果之一。
作者简介:李慧(ORCID:0000-0002-3468-5170),副教授,博士;玄洪升(ORCID:0000-0002-1837-5105),硕士研究生,通讯作者,E-mail:summer.ai.xuan@gmail.com。
收稿日期:2019-11-07 修回日期:2020-02-27 本文起止页码:96-107 本文责任编辑:易飞

析算法识别技术主题。O. Kwon 等[8]构建了专利引文耦合网络和共引网络,综合分析专利分布情况从而识别技术主题。张欣等[9]将改进的 PageRank 算法与专利的被引次数和专利年龄结合,并将其应用到 OLED 领域中来识别核心专利。随着文本聚类、LDA 主题模型[10]和社区识别[11－12]等自然语言处理技术的发展,基于内容特征的技术主题识别方法也逐渐受到学者的重视。C. Hayoung 等[13]提出识别专利潜在技术创新主题的算法,利用增强现实专利的摘要内容,识别出具有潜在创新价值的技术主题。伊惠芳等[14]结合 LDA 模型和战略坐标图方法进行专利文本内容分析,识别出技术主题及其结构特征。范宇等[15]提出了应用于专利内容聚类的主题模型和聚类算法,将潜在狄利克雷分布(LDA) 主题模型和 OPTICS 算法相结合进行核心技术主题分析。综合现有研究分析发现,虽然基于专利引文特征的识别方法能够较为有效地识别出领域技术主题,但由于引文分析存在引文时滞性,所以,识别出的技术主题在时效性、准确性方面存在一定的缺陷。再者,基于专利文本内容相比基于引文特征的方法具有一定的优势(不存在引文时滞性),但同样存在一定的不足,如从专利标题、摘要等文本内容中挖掘技术主题,仅仅从自然语言处理的角度进行考量,并没有考虑技术主题需要具备的经济和技术属性。

综上所述,针对目前利用专利数据进行技术主题挖掘的不足,本文提出融合多属性的量化方法,快速且有效地挖掘出领域内多个技术创新主题,其中,技术创新主题是指可以发展或改进的广义技术主题[13]。主要创新之处在于:①避免专利引文分析的时滞性。使用权利要求数与 IPC 分类数[16]替代引文量,可以随机选取大量的领域专利作为语料挖掘技术主题,最大程度地规避人为因素影响最终结果。②弥补技术主题在经济和技术属性方面的缺失。研究表明有价值的专利表现为专利权利要求的数量多而且技术覆盖范围广[17],其中专利涉及的 IPC 分类越多,则说明该专利涉及的技术领域越广。因此,本文引入专利价值评价指标[16]中的权利要求数和 IPC 分类。③融合多属性构建量化指标体系。综合研究 LDA 概率分布矩阵与专利价值评价指标,多维度定义量化指标,构建识别技术创新主题的量化指标体系。

## 2　研究设计

本文的研究设计前后分为三个部分:①技术特征词向量化;②量化指标体系构建;③技术创新主题挖掘,其中②量化指标体系构建是本文理论研究的核心,同时也是主要创新所在。研究设计的具体过程如图 1 所示:

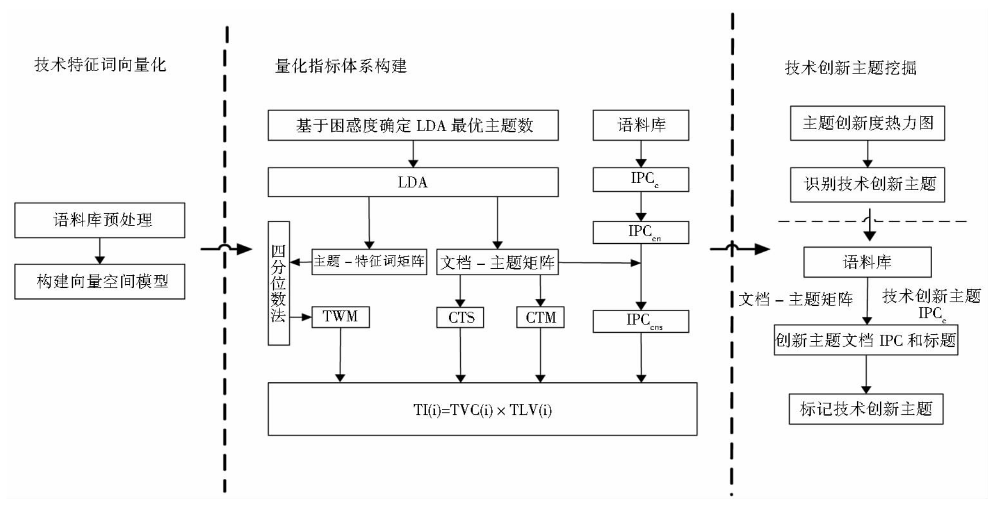



图 1　研究设计

### 2.1　技术特征词向量化

技术特征词向量化主要分为两部分:数据预处理和构建向量空间模型。

(1)数据预处理。首先对语料进行分词,然后去除停用词、词干还原,最后去除标点符号、特殊符号和数字。在词频矩阵中仍会出现一些噪声词汇,如:method、system、action 等,通过编写程序去除这些噪声词汇。

(2)构建向量空间模型[18]。首先根据确定的技术特征词数量,将预处理后的语料转换为词频(TF)矩阵[19];然后将词频(TF)矩阵转换成逆文本词频(IDF)矩阵[19];最后将 TF 与 IDF 矩阵相乘生成 TF-IDF 矩阵。

## 2.2 量化指标体系构建

量化指标体系构建是本文研究成果的重点所在，通过构建量化指标体系，识别出技术创新主题[13]，不再是仅基于自然语言处理属性的技术主题，还将经济和技术属性考量在内。

主要分为两个部分：构建 LDA 主题模型[20]和构建量化指标体系。

### 2.2.1 构建 LDA 主题模型

首先使用基于困惑度[20]的方法确定最优主题数，目前普遍认为应用 LDA 的最大问题是无法确定最优主题数目[21]。本文采用困惑度方法确定 LDA 主题模型的最优主题数。将数据集分为训练集与测试集，使用 TF-IDF 对数据集加权处理，利用加权后的训练集构建 LDA 模型，由于 LDA 在专利文本分析方面的优势[22-24]，我们将基于 LDA 概率主题建模生成专利文档－主题和主题－特征词的概率分布矩阵。等模型训练结束后，将测试集作为语料计算 LDA 模型在不同主题下的困惑度，最终选取困惑度最小时的主题数作为模型的最优主题数；然后正式构建 LDA 主题模型，此时加权语料集和最优主题数两项重要构建元素准备完成，构建 LDA 主题模型，最终生成文档－主题矩阵和主题－特征词矩阵。

### 2.2.2 构建量化指标体系

为了使技术主题不仅具有自然语言处理的属性而且具有经济和技术属性，运用量化思想处理 LDA 输出的概率分布矩阵、权利要求数和 IPC 分类数，然后根据相关理论研究成果，定义量化指标，构建识别技术创新主题的三级量化指标体系，如图 2 所示。

构建量化指标体系分为三部分：①Ⅲ级量化指标定义；②Ⅱ级量化指标定义；③Ⅰ级量化指标定义。

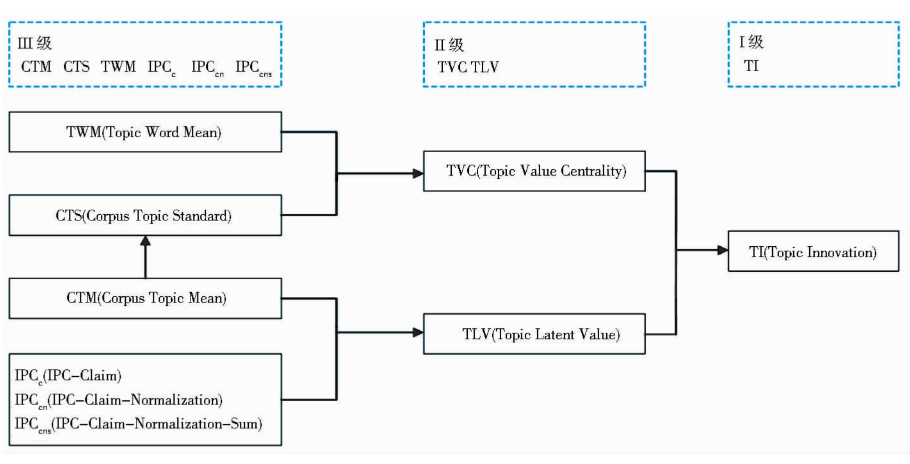



图 2 量化指标体系



（1）Ⅲ级量化指标定义。LDA 主题模型构建完毕，生成两个概率分布矩阵（文档－技术主题、技术主题－特征词）。基于此，我们将定义具有自然语言处理属性的量化指标（CTM、CTS、TWM），并基于 IPC 分类数和权利要求数定义具有经济和技术属性的量化指标（$IPC_c$、$IPC_{cn}$、$IPC_{cns}$）。

CTM（Corpus Topic Mean）表示在语料库范围内技术主题概率的均值[25]，计算公式如式（1）所示。CTM 表示技术主题在当前语料库内技术价值[16]的大小，CTM 值越大，表示技术主题在当前语料库内所具有的技术价值越大，反之亦然。

$$CTM(j) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} t_{ij} \qquad 式(1)$$

其中，N 表示语料库内专利文档的数量；M 表示主题数量；$t_{ij}$表示主题 j 在第 i 篇专利文档的概率值。

CTS（Corpus Topic Standard）表示在语料库范围内技术主题概率的标准差[25]，计算公式如式（2）所示。CTS 表示技术主题在当前语料库内技术价值[16]的稳定性，稳定性衡量该技术主题在当前语料库内技术价值的离散程度。CTS 值越小，表示技术主题在当前语料库内的技术价值越稳定，反之亦然。

$$CTS(j) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}(t_{ij} - CTM_j)} \qquad 式(2)$$

其中，N 表示语料库内专利文档的数量；M 表示主题数量；$t_{ij}$ 表示主题 j 在第 i 篇专利文档的概率值；$CTM_j$ 表示主题 j 在语料库中的均值。

TWM（Topic Word Mean）表示技术主题的特征词概率的均值[25]，计算公式如式（3）所示。在计算 TWM 时，为了选择对技术主题解释能力强的特征词，引入四分位数法[26]将每个技术主题下的特征词按照概率值降序排序，选择其中前四分之一的特征词计算 TWM 的值，间接地优化了主题－特征词概率分布矩阵。TWM

表示技术主题被解释程度的大小，TWM 值越大，表示技术主题被解释的越充分，即技术主题当前具有的技术价值越具有说服力，反之亦然。

$$TWM(j)=\frac{1}{K}\sum_{i=1}^{K}\sum_{j=1}^{M}t_{ij} \quad 式(3)$$

其中，K 表示特征词的数量；M 表示主题数量；$t_{ij}$ 表示主题 j 的第 i 个特征词的概率。TWM(j)表示主题 j 的特征词概率均值。

IPC 分类数和权利要求数是专利技术价值的评价指标[16]，与专利评价指标的被引文数相比，前面两者的数量不会随着时间变化而改变，后者则会随着时间动态增长。IPC 分类数表示专利技术的覆盖范围，研究表明[16]：专利的 IPC 分类数越大，其技术价值越高，产生的经济效益也越大；权利要求数表示专利技术被保护宽度，研究表明权利要求数与专利技术价值有很好的相关性。专利的被引文量在评估专利技术价值方面有不可替代的作用[27]，但在评估新近发表的专利价值方面，其评估作用远不如 IPC 分类数和权利要求数，这尤其适应本文研究专利数据的时间特征。原始的专利文献中有 IPC 分类号和权利要求两个条目，运用统计方法，计算出每篇专利文献中 IPC 分类号的数量和权利要求的数量。利用 IPC 分类数和权利要求数不仅可以评估其技术价值，而且可以评估其潜在技术价值。运用量化方法的思想，将两者分别乘以调节系数(α、β)后相加，相加后的数值表示专利具有的潜在技术价值。

$IPC_c$(IPC-Claim)表示专利的潜在技术价值，计算公式如式(4)所示。IPCc 表示专利具有的潜在技术价值，IPCc 值越大，表示专利具有的潜在技术价值越大，反之亦然。权利要求数与 IPC 分类数是专利文本的两项独立的指标，且权利要求数往往大于 IPC 分类数，又因不同的专利技术领域内两者的差异有不同，所以设置调节系数 α 和 β[28] 使两者对整体的贡献度相同，调节系数的设置取决于当前语料库内数据，计算公式如式(5)、(6)所示。

$$IPC_c(i)=\alpha N_Claim_i+\beta N_IPC_i(i<=N \quad 0<\alpha<=1 \quad \beta>=1) \quad 式(4)$$

$$\alpha=\frac{1}{2}\left(1+\frac{\sum_{j=1}^{N}IPC_j}{\sum_{i=1}^{N}Claim_i}\right) \quad 式(5)$$

$$\beta=\frac{1}{2}\left(1+\frac{\sum_{i=1}^{N}Claim_i}{\sum_{j=1}^{N}IPC_j}\right) \quad 式(6)$$

其中，N 表示语料库专利文档的数量；$N_Claim_i$ 表示第 i 篇专利文档中权利要求的数量；$N_IPC_i$ 表示第 i 篇专利文档中 IPC 分类号的数量；$\sum_{i=1}^{N}Claim_i$ 表示语料库内权利要求数的总和；$\sum_{j=1}^{N}IPC_j$ 表示语料库内 IPC 分类数的总和。

IPCcn (IPC-Claim-Normalization) 是将 IPCc 归一化，如式(7)所示。由于 IPCc 的数值较大会影响到后面技术主题潜在价值度(TLC)的定义和科学计算结果，经实验和讨论后决定将 IPCc 进行离差标准化[29]处理，经试验检测，计算结果符合实验的预期效果。

IPCcn 仍表示专利具有的潜在技术价值。

$$IPC_{cn}(i)=\frac{IPC_{c_i}-IPC_{c_{min}}}{IPC_{c_{max}}-IPC_{c_{min}}}(i<=N) \quad 式(7)$$

其中，$IPC_{c_i}$ 表示 IPCc 中的第 i 个值；$IPC_{c_{min}}$ 表示 IPCc 中最小的值；$IPC_{c_{max}}$ 表示 IPCc 中最大的值。

$IPC_{cns}$ (IPC-Claim-Normalization-Sum) 是技术主题的 $IPC_{cn}$ 之和，如公式(8)所示。由图 1 研究设计图可以观察到，文档－主题矩阵指向连接 $IPC_{cn}$ 和 $IPC_{cns}$ 的有向线段的中部。在定义 $IPC_{cns}$ 时，为了解决表示技术主题潜在技术价值不同的难题，通过实验为文档－主题矩阵设置合适的阈值，筛选出每篇文档中典型的技术主题，新的文档－主题矩阵中使用数字 1 表示技术主题出现在当前专利文档，数字 0 表示技术主题未出现在当前专利文档，这样就可以建立新文档－主题矩阵与 $IPC_{cn}$ 的映射关系，匹配出每个技术主题各自对应的 $IPC_{cn}$，从而解决了如何表示技术主题潜在技术价值的问题，完成对 $IPC_{cns}$ 的定义和计算。$IPC_{cns}$ 表示技术主题所具有的专利潜在技术价值之和，$IPC_{cns}$ 的值越大，表示技术主题具有潜在技术价值的专利越多，反之亦然。

$$IPC_{cns}(j)=\sum_{i=1}^{P}\sum_{j=1}^{M}IPC_{cn_{ij}} \quad 式(8)$$

其中，P 表示属于每个技术主题的专利文档的数量，每个技术主题的专利文档的数量不同；M 表示技术主题数量；$IPC_{cns}$(j)表示技术主题 j 的 $IPC_{cn}$ 累加之和。

(2) Ⅱ级量化指标定义。TVC (Topic Value Centrality) 是技术主题的 CTS (Corpus Topic Standard) 的倒数与 TWM (Topic Word Mean) 的乘积，如式(9)所示。TVC 表示技术主题中心性的强弱，即技术主题在当前阶段所具有的技术价值。TVC 值越大，表示技术主题在当前所具有的技术价值越大，反之亦然。

$$TVC(j)=\frac{1}{CTS_j}\times TWM_j(i<=M) \quad 式(9)$$

M 表示主题的数量；$CTS_j$ 表示主题 j 的语料库主

题概率标准差值；$TWM_j$ 表示主题 j 的特征词概率平均值。

TLV（Topic Latent Value）是技术主题的 CTM（Corpus Topic Mean）与 $IPC_{cns}$ 的乘积，计算公式如式（10）所示。TLV 表示技术主题潜在技术价值的大小，即技术主题在未来阶段所具有的技术价值。TLV 值越大，表示技术主题在未来所具有的技术价值越大，反之亦然。

$$TLV(j) = CTM_j \times IPC_{cns_j} (j < = M) \qquad 式(10)$$

M 表示主题的数量；$CTM_j$ 表示主题 j 在语料库内的均值；$IPC_{cns_j}$ 表示主题 j 的 $IPC_{cn}$ 之和。

（3）Ⅰ级量化指标定义。TI（Topic Innovation）是 TVC（Topic Value Centrality）与 TLV（Topic Latent Value）的乘积，同时也是融合多属性来识别技术创新主题的量化指标，如公式（11）所示。TI 表示技术主题创新性的强弱，即技术主题所具有的创新价值。TI 值越大，表示技术主题的创新价值越大，反之亦然。

$$TI(j) = TVC_j \times TLV_j (j < = M) \qquad 式(11)$$

M 表示主题的数量；$TVC_j$ 表示主题 j 的技术主题价值中心度值；$TIV_j$ 表示主题 j 的技术主题创新价值度值。

### 2.3　技术创新主题挖掘

#### 2.3.1　技术创新主题识别

量化指标体系构建完毕，技术创新主题可以通过技术主题创新度（TI）识别出来，但是单纯的数值对于结果的呈现效果并不佳，借助知识图谱呈现主题创新度的结果，可以直观地识别出技术创新主题。

#### 2.3.2　技术创新主题标记

技术创新主题标记阶段是本文研究的汇聚阶段，前面 4 个阶段都在为这个阶段做准备。技术创新主题虽被识别出来，但每个技术创新主题并没有一个恰当的标记，这个阶段的任务就是利用前面 4 个阶段已有的数据结果标记创新主题。根据 $IPC_c$ 数对属于每个创新主题的专利设置不同的阈值，挑选出合适数量的专利文档。经过讨论，决定使用主题下专利的 IPC 分类说明和特征词定义技术创新主题，但在实验阶段发现创新主题的某些特征词专业性不强，经过反复讨论和实验，确定将创新专利文档的标题分割去重，然后从中挑选合适的词汇替换创新主题中专业性较弱的特征词。最终根据 IPC 分类说明和优化后的特征词完成对技术创新主题的标记。

## 3　实验验证

### 3.1　获取数据和预处理

实验验证语料选择芯片领域的专利文献，从专利数据库 Total Patent 下载 2014 – 2018 年芯片领域的英文专利文献，共计 9 197 篇；检索表达式为 Ti：（integrated circuit OR microcircuit OR microchip OR chip）；下载的专利文档条目包括标题、摘要、IPC 分类号、权利要求。

构建 LDA 主题模型的语料使用芯片专利文档条目中的摘要，利用 Python 的自然语言工具包 – NLTK[30] 完成对摘要的预处理。

### 3.2　构建主题模型

#### 3.2.1　确定最优主题数

在构建 LDA 主题模型之前，除了需要准备预处理过的语料，还要确定最优主题数，本文使用工具包 – sklearn[31] 计算 0 至 100 之间主题数的困惑度值，完成最优主题数的确定，如图 3 所示：

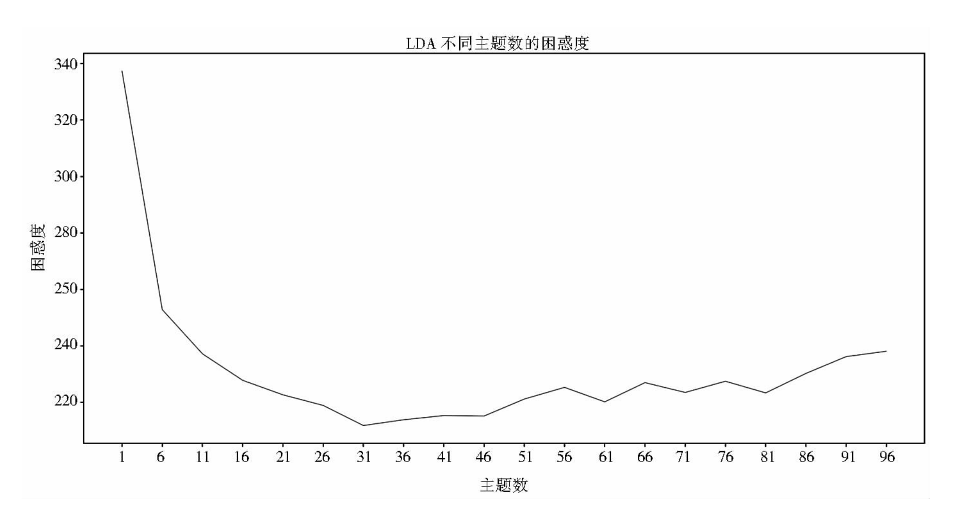



图 3　主题数困惑度

由图3可以看出,主题数在31附近时的困惑度最小,因此确定31为LDA主题模型的最优主题数。

### 3.2.2 构建LDA主题模型

向量空间模型的构建使用工具包-sklearn[31],设置特征词数为2 000个,然后生成词频(TF)矩阵、逆文本(IDF)矩阵、TF-IDF矩阵,三个矩阵的行列数均为9 197和2 000。读取TF-IDF矩阵,将LDA的主题数设置为31,迭代次数设置为100,LDA的初始化信息如表1所示。LDA模型的构建使用LDA工具包[32],生成文档-主题(Doc-Topic)概率分布矩阵和主题-特征词(Topic-Term)概率分布矩阵,矩阵的局部如表2、表3所示:

表1 LDA初始化信息

| 序号 | 初始化条目 | 数量 |
|---|---|---|
| 1 | 专利文档数 | 9 197 |
| 2 | 特征词数 | 2 000 |
| 3 | 分词数 | 1 157 761 692 |
| 4 | 主题数 | 31 |
| 5 | 迭代次数 | 100 |

表2 文档-主题(Doc-Topic)概率分布矩阵局部

| | Topic-1 | Topic-2 | Topic-3 | Topic-4 | Topic-5 |
|---|---|---|---|---|---|
| Doc-1 | 9.37E-07 | 9.37E-07 | 12.7E-02 | 9.37E-07 | 3.60E-02 |
| Doc-2 | 1.39E-06 | 1.39E-06 | 1.39E-06 | 1.39E-06 | 1.39E-06 |
| Doc-3 | 8.09E-07 | 8.09E-07 | 4.09E-02 | 8.09E-07 | 1.61E-01 |
| Doc-4 | 1.09E-06 | 1.09E-06 | 1.09E-06 | 1.09E-06 | 1.60E-01 |
| Doc-5 | 5.27E-07 | 5.27E-07 | 5.27E-07 | 9.15E-01 | 5.27E-07 |

表3 主题-特征词(Topic-Term)概率分布矩阵局部

| | 3d | 3dic | 3dics | affix | agent |
|---|---|---|---|---|---|
| Topic-1 | 3.27E-10 | 3.27E-10 | 3.27E-10 | 3.27E-10 | 3.27E-10 |
| Topic-2 | 1.78E-04 | 3.60E-10 | 3.60E-10 | 3.60E-10 | 3.60E-10 |
| Topic-3 | 3.34E-07 | 1.75E-10 | 1.75E-10 | 1.81E-04 | 1.75E-10 |
| Topic-4 | 3.41E-07 | 2.13E-10 | 2.13E-10 | 4.92E-04 | 2.13E-10 |
| Topic-5 | 1.25E-10 | 1.25E-10 | 1.25E-10 | 1.25E-10 | 1.25E-10 |

## 3.3 计算量化指标

### 3.3.1 基于LDA概率分布矩阵计算量化指标

(1)计算CTM(Corpus Topic Mean)和CTS(Corpus Topic Standard)。利用文档-主题矩阵分别计算每个芯片技术主题在语料库范围内概率的平均值(式(1))和标准差(式(2))。最终的计算结果保留两位小数,如表4、表5和图4所示:

表4 CTM

| 序号 | 技术主题 | CTM | 序号 | 技术主题 | CTM |
|---|---|---|---|---|---|
| 1 | Topic-1 | 0.03 | 17 | Topic-17 | 0.02 |
| 2 | Topic-2 | 0.02 | 18 | Topic-18 | 0.02 |
| 3 | Topic-3 | 0.05 | 19 | Topic-19 | 0.02 |
| 4 | Topic-4 | 0.04 | 20 | Topic-20 | 0.02 |
| 5 | Topic-5 | 0.07 | 21 | Topic-21 | 0.03 |
| 6 | Topic-6 | 0.03 | 22 | Topic-22 | 0.03 |
| 7 | Topic-7 | 0.02 | 23 | Topic-23 | 0.02 |
| 8 | Topic-8 | 0.04 | 24 | Topic-24 | 0.01 |
| 9 | Topic-9 | 0.02 | 25 | Topic-25 | 0.04 |
| 10 | Topic-10 | 0.03 | 26 | Topic-26 | 0.03 |
| 11 | Topic-11 | 0.02 | 27 | Topic-27 | 0.06 |
| 12 | Topic-12 | 0.03 | 28 | Topic-28 | 0.05 |
| 13 | Topic-13 | 0.03 | 29 | Topic-29 | 0.03 |
| 14 | Topic-14 | 0.04 | 30 | Topic-30 | 0.02 |
| 15 | Topic-15 | 0.04 | 31 | Topic-31 | 0.02 |
| 16 | Topic-16 | 0.03 | | | |

表5 CTS

| 序号 | 技术主题 | CTS | 序号 | 技术主题 | CTS |
|---|---|---|---|---|---|
| 1 | Topic-1 | 0.07 | 17 | Topic-17 | 0.07 |
| 2 | Topic-2 | 0.09 | 18 | Topic-18 | 0.09 |
| 3 | Topic-3 | 0.12 | 19 | Topic-19 | 0.08 |
| 4 | Topic-4 | 0.12 | 20 | Topic-20 | 0.1 |
| 5 | Topic-5 | 0.12 | 21 | Topic-21 | 0.03 |
| 6 | Topic-6 | 0.11 | 22 | Topic-22 | 0.1 |
| 7 | Topic-7 | 0.08 | 23 | Topic-23 | 0.09 |
| 8 | Topic-8 | 0.09 | 24 | Topic-24 | 0.07 |
| 9 | Topic-9 | 0.08 | 25 | Topic-25 | 0.1 |
| 10 | Topic-10 | 0.09 | 26 | Topic-26 | 0.08 |
| 11 | Topic-11 | 0.08 | 27 | Topic-27 | 0.11 |
| 12 | Topic-12 | 0.08 | 28 | Topic-28 | 0.12 |
| 13 | Topic-13 | 0.09 | 29 | Topic-29 | 0.09 |
| 14 | Topic-14 | 0.12 | 30 | Topic-30 | 0.09 |
| 15 | Topic-15 | 0.09 | 31 | Topic-31 | 0.09 |
| 16 | Topic-16 | 0.08 | | | |

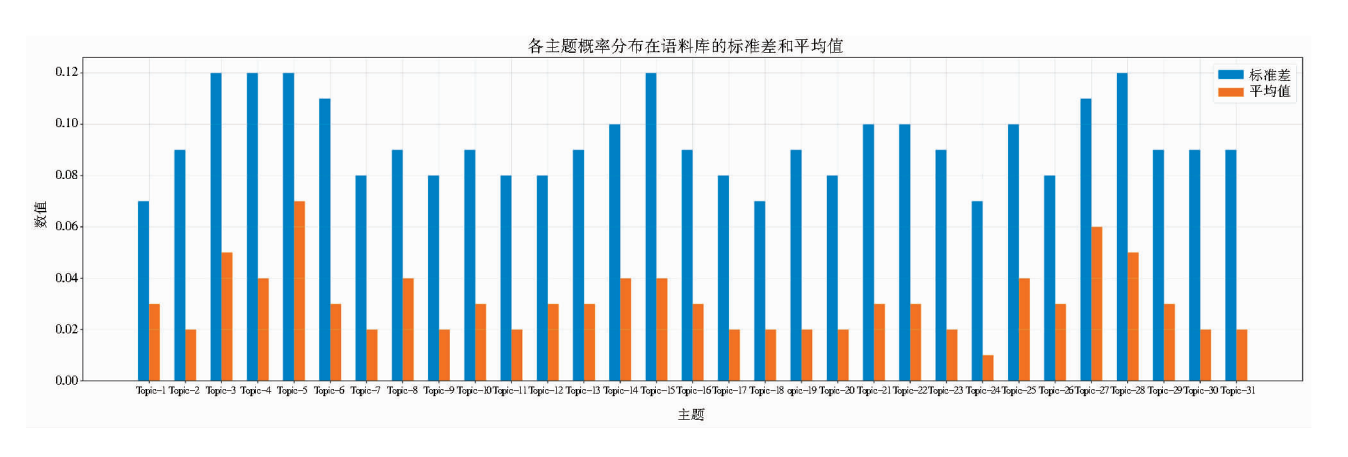



图 4 CTM 和 CTS

（2）计算 TWM（Topic Word Mean）。利用主题－特征词矩阵计算每个芯片技术主题的特征词概率平均值（式（3））。首先使用四分位数法[26]选取各技术主题的有效特征词，每个技术主题选取了 32 个最能表达主题的特征词，如表 6 所示，然后计算各主题下特征词概率平均值（TWM），最终的计算结果保留两位小数，如表 7 所示。

表 6 基于四分位数法的主题－特征词概率分布矩阵局部

| 序号 | Topic－1 | 概率值降序排序 | Topic－2 | 概率值降序排序 | Topic－3 | 概率值降序排序 | Topic－4 | 概率值降序排序 |
|---|---|---|---|---|---|---|---|---|
| 1 | device | 0.289 200 135 | wafer | 0.209 738 975 | die | 0.379 672 634 | signal | 0.171 639 651 |
| 2 | couple | 0.108 120 538 | unit | 0.099 284 345 | package | 0.194 450 518 | output | 0.061 679 928 |
| 3 | vias | 0.071 547 453 | support | 0.098 626 190 | ic | 0.138 210 838 | sensor | 0.057 841 045 |
| 4 | interface | 0.059 826 328 | sense | 0.063 709 929 | polymer | 0.029 836 326 | input | 0.038 605 892 |
| 5 | pin | 0.053 534 709 | underfill | 0.026 204 514 | encapsulant | 0.026 334 772 | image | 0.037 023 626 |
| 7 | external | 0.042 067 096 | ppi | 0.021 817 573 | attach | 0.024 680 234 | port | 0.032 833 949 |

表 7 TWM

| 序号 | 技术主题 | TWM | 序号 | 技术主题 | TWM |
|---|---|---|---|---|---|
| 1 | Topic-1 | 0.28 | 17 | Topic-17 | 0.25 |
| 2 | Topic-2 | 0.25 | 18 | Topic-18 | 0.28 |
| 3 | Topic-3 | 0.29 | 19 | Topic-19 | 0.24 |
| 4 | Topic-4 | 0.22 | 20 | Topic-20 | 0.28 |
| 5 | Topic-5 | 0.31 | 21 | Topic-21 | 0.28 |
| 6 | Topic-6 | 0.28 | 22 | Topic-22 | 0.23 |
| 7 | Topic-7 | 0.26 | 23 | Topic-23 | 0.24 |
| 8 | Topic-8 | 0.30 | 24 | Topic-24 | 0.27 |
| 9 | Topic-9 | 0.27 | 25 | Topic-25 | 0.27 |
| 10 | Topic-10 | 0.22 | 26 | Topic-26 | 0.24 |
| 11 | Topic-11 | 0.28 | 27 | Topic-27 | 0.30 |
| 12 | Topic-12 | 0.28 | 28 | Topic-28 | 0.30 |
| 13 | Topic-13 | 0.29 | 29 | Topic-29 | 0.22 |
| 14 | Topic-14 | 0.30 | 30 | Topic-30 | 0.24 |
| 15 | Topic-15 | 0.18 | 31 | Topic-31 | 0.27 |
| 16 | Topic-16 | 0.29 | | | |

### 3.3.2 基于专利文献评价指标计算量化指标

（1）计算 IPCc（IPC-Claim）和 IPCcn（IPC-Claim-Normalization）。读取芯片专利语料库的 IPC 和权利要求条目数据，然后将 IPC 分类号和权利要求按各自条目组织数据的特点分割成单一 IPC 和权利要求，运用统计学知识统计两个条目中分割后的数量，替代原 IPC 和权利要求条目，更新后的 IPC 和权利要求条目如表 8 所示。最后计算每篇专利的 IPCc（式（4）），在芯片语料库新建 IPCc 条目，如表 8 中的新增条目 IPCc 所示。

（2）计算 IPCcn（IPC-Claim-Normalization）。使用离差标准化方法[29]将 IPCc 条目的数据归一化处理后，利用结果数据在芯片语料库新建 IPCcn 条目，计算（式（7））结果如表 8 中的新增条目 IPCcn 所示。

### 3.4 识别技术创新主题

这一节的实验共分为两部分，计算技术主题价值中心度（TVC）、技术主题潜在价值度（TLV）、主题创新度（TI），并可视化识别技术创新主题。

表8　芯片专利语料库新建 IPCc 和 IPCcn 条目局部

| 序号 | 摘要 | 权利要求 | IPC | 标题 | IPCc | IPCcn |
|---|---|---|---|---|---|---|
| 1 | A package is formed by | 20 | 4 | Stress relieved th | 24 | 0.17 |
| 2 | An integrated circuit in | 18 | 5 | Integrated circuits | 23 | 0.17 |
| 3 | A flexible package may | 20 | 6 | FLEXIBLE PAC | 26 | 0.19 |
| 4 | Electronic module, wh | 20 | 7 | CHIP ASSEMB | 27 | 0.20 |
| 5 | Methods and structures | 20 | 5 | Packaging identic | 25 | 0.18 |
| 6 | An embodiment of a m | 26 | 3 | Indexing of electr | 29 | 0.21 |
| 7 | A power and ground sh | 9 | 8 | Power and groun | 17 | 0.12 |
| 8 | Metal pillars are placed | 20 | 31 | Solder bump plac | 51 | 0.40 |
| 9 | Chip package structures | 20 | 9 | Method for formin | 29 | 0.21 |
| 10 | The fabrication of an el | 20 | 4 | Electronic docum | 24 | 0.17 |

(1)参照式(9)、(10)、(11)分别计算 TVC、TLV 和 TI,计算结果保留两位小数,如表9所示:

表9　技术主题创新度

| 序号 | 技术主题 | 技术主题价值中心度(TVC) | 技术主题潜在价值度(TLV) | 技术主题创新度(TI) |
|---|---|---|---|---|
| 1 | Topic-1 | 20.00 | 17.59 | 351.80 |
| 2 | Topic-2 | 13.89 | 6.17 | 85.70 |
| 3 | Topic-3 | 12.08 | 28.81 | 348.02 |
| 4 | Topic-4 | 9.17 | 14.52 | 133.15 |
| 5 | Topic-5 | 12.92 | 61.11 | 789.54 |
| 6 | Topic-6 | 12.73 | 11.96 | 152.25 |
| 7 | Topic-7 | 16.25 | 5.29 | 85.96 |
| 8 | Topic-8 | 16.67 | 21.75 | 362.57 |
| 9 | Topic-9 | 16.88 | 5.93 | 100.10 |
| 10 | Topic-10 | 12.22 | 9.23 | 112.79 |
| 11 | Topic-11 | 17.50 | 7.70 | 134.75 |
| 12 | Topic-12 | 17.50 | 11.20 | 196.00 |
| 13 | Topic-13 | 16.11 | 10.81 | 174.15 |
| 14 | Topic-14 | 15.00 | 23.30 | 349.50 |
| 15 | Topic-15 | 7.50 | 14.23 | 106.72 |
| 16 | Topic-16 | 16.11 | 14.14 | 227.80 |
| 17 | Topic-17 | 15.62 | 6.91 | 107.93 |
| 18 | Topic-18 | 20.00 | 8.91 | 178.20 |
| 19 | Topic-19 | 13.33 | 6.53 | 87.04 |
| 20 | Topic-20 | 17.50 | 6.62 | 115.85 |
| 21 | Topic-21 | 14.00 | 10.67 | 149.38 |
| 22 | Topic-22 | 11.50 | 9.37 | 107.76 |
| 23 | Topic-23 | 13.33 | 5.13 | 68.38 |
| 24 | Topic-24 | 19.29 | 1.70 | 32.79 |
| 25 | Topic-25 | 13.50 | 28.92 | 390.42 |
| 26 | Topic-26 | 15.00 | 13.26 | 198.90 |
| 27 | Topic-27 | 13.64 | 45.84 | 625.26 |
| 28 | Topic-28 | 12.50 | 29.90 | 373.75 |
| 29 | Topic-29 | 12.22 | 9.99 | 122.08 |
| 30 | Topic-30 | 13.33 | 6.35 | 84.65 |
| 31 | Topic-31 | 15.00 | 4.78 | 71.70 |



(2)可视化识别技术创新主题。本文研究的中心思想是通过提出一种融合多属性的量化方法,快速有效地挖掘出多个技术创新主题,具体思路就是首先利用 LDA 挖掘出技术主题,其次融合多属性从这些技术主题中挖掘出几个可能最具创新价值的技术主题,然后通过邀请领域专家对挖掘出的技术主题做一个专业的评判,如果专家的评分表明技术主题有价值,那么本文的研究工作就是有意义的;相反,如果专家的评分表明技术主题没有价值,那么本文提出的方法就是错误的。通过 TI 值大小选取5个技术主题,符合本文的研究目的——快速且有效地挖掘多个可能最具创新价值的技术主题。

可视化时借助热力图呈现 TVC、TLV 和 TI 的计算结果。实验过程中,发现由于 TI 的值相对较大,导致热力图中技术主题的 TVC 和 TLV 区别不明显。重复几次实验验证,将 TI 值缩小5倍后,热力图效果最佳,可视化结果见图5。

通过直观的热力图和 TI 值排序表(见表10),选择 TI 值前5的技术主题作为技术创新主题,分别是 Topic-5、Topic-27、Topic-25、Topic-28 和 Topic-8。

## 3.5　技术创新主题的标记

(1)数据准备。读取有新建条目 IPCc 的芯片语料库,然后读取文档－主题矩阵。

(2)数据处理。其一,提取技术创新主题专利文档。首先将5个技术创新主题在 LDA 生成的文档－主题矩阵中对应的部分构成一个新的文档－创新主题矩阵,建立语料库与新矩阵的映射关系,提取出符合条件的专利文档。由于符合条件的专利文档数量较大,设置不同阈值选取专利文档,依次将新矩阵中的概率值保留两位小数、一位小数和不保留小数,以创新主题 Topic-5 为例提取到专利文档数量分别是 5 284、3 256 和 122,其他4个主题呈现相同的递减规律。经过讨论

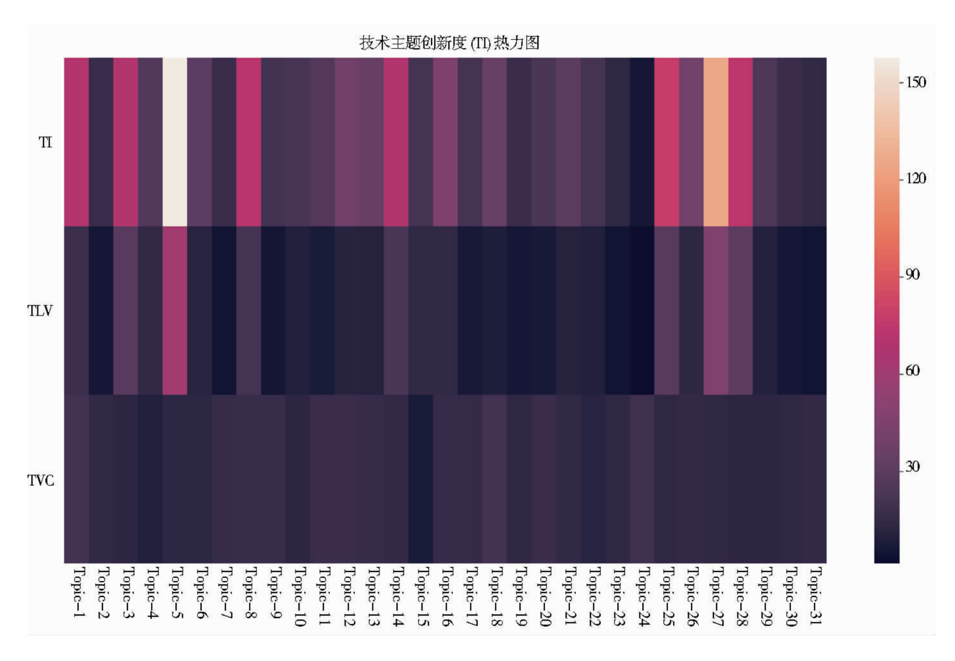



图 5　技术主题创新度热力图

表 10　技术主题创新度排序

| 序号 | 技术主题 | 技术主题创新度(TI) | 序号 | 技术主题 | 技术主题创新度(TI) |
|---|---|---|---|---|---|
| 1 | Topic-5 | 157.66 | 17 | Topic-4 | 26.57 |
| 2 | Topic-27 | 125.14 | 18 | Topic-29 | 24.38 |
| 3 | Topic-25 | 78.08 | 19 | Topic-20 | 23.17 |
| 4 | Topic-28 | 74.75 | 20 | Topic-10 | 22.52 |
| 5 | Topic-8 | 72.43 | 21 | Topic-17 | 21.56 |
| 6 | Topic-1 | 70.36 | 22 | Topic-22 | 21.55 |
| 7 | Topic-14 | 69.90 | 23 | Topic-15 | 21.34 |
| 8 | Topic-3 | 69.72 | 24 | Topic-9 | 20.04 |
| 9 | Topic-16 | 45.53 | 25 | Topic-19 | 17.44 |
| 10 | Topic-26 | 39.78 | 26 | Topic-7 | 17.19 |
| 11 | Topic-12 | 39.20 | 27 | Topic-2 | 17.15 |
| 12 | Topic-18 | 35.64 | 28 | Topic-30 | 16.95 |
| 13 | Topic-13 | 34.81 | 29 | Topic-31 | 14.34 |
| 14 | Topic-6 | 30.50 | 30 | Topic-23 | 13.70 |
| 15 | Topic-21 | 29.88 | 31 | Topic-24 | 6.56 |
| 16 | Topic-11 | 26.95 | | | |

确定使用不保留小数的矩阵生成布尔索引与语料库映射提取合理数量的专利文档，Topic-5、Topic-8、Topic-25、Topic-27、Topic-28 这 5 个创新主题对应文档数量分别是 122、87、121、134、165。

其二，将 IPC 和标题分割、去重。每个技术创新主题对应的专利文档已经明确，分别提取各技术创新主题的专利文档的 IPC 和标题条目内容，根据各自条目的内容特点进行分割然后去重操作，去重后统计每个主题下 IPC 的数量，如表 11 所示，数量太大，不利于创新主题的标记。

其三，筛选技术创新主题的专利文档。由于 IPC 的数量问题，导致无法有效标记技术创新主题，经过实验和讨论，解决这个问题的方法是首先将创新主题的专利文档按照 IPCc 的数量排序，并按 IPCc 数分组统计专利文档的数量和 IPC 的数量，然后再将 Topic-5、Topic-8、Topic-25、Topic-27、Topic-28 依次分别取 IPCc 的数量大于 32、29、29、31 和 30 的分组，最后统计筛选后的 5 个创新主题的专利文档数和 IPC 数的结果，如表 12 所示：

表 11　技术创新主题的 IPC 去重后的数量

| 序号 | 技术创新主题 | IPC 数 |
|---|---|---|
| 1 | Topic-5 | 123 |
| 2 | Topic-8 | 100 |
| 3 | Topic-25 | 169 |
| 4 | Topic-27 | 97 |
| 5 | Topic-28 | 130 |

表 12　统计标记技术创新主题的专利文档数量和 IPC 数量

| 序号 | 技术创新主题 | 专利文档数 | IPC 数 |
|---|---|---|---|
| 1 | Topic-5 | 10 | 31 |
| 2 | Topic-8 | 13 | 27 |
| 3 | Topic-25 | 10 | 29 |
| 4 | Topic-27 | 11 | 40 |
| 5 | Topic-28 | 11 | 44 |

(3)标记结果。技术创新主题的标记根据优化后的专利 IPC 说明[33]和创新特征词，其中创新特征词的确定是基于专利文档标题条目内容的分词结果和四分位数法优化后的主题－特征词矩阵中的特征词，每个

创新主题都由30个创新特征词标记,将Topic-5、Topic-27、Topic-25、Topic-28、Topic-8按照TI值大小依次重新命名为C-Topic-1、C-Topic-2、C-Topic-3、C-Topic-4、C-Topic-5,根据IPC说明和特征词完成对技术创新主题的标记。创新主题的标记结果如表13所示:

表13　创新主题标记

| 技术创新主题 | 标记结果 |
|---|---|
| C-Topic-1 | 围绕半导体衬底的封装技术,包括管芯相对于衬底的位置以及衬底本身的特殊构造,如凹槽等。 |
| C-Topic-2 | 接触垫在芯片封装中的应用。 |
| C-Topic-3 | 集成电路或系统中模块的功能配置及互连。 |
| C-Topic-4 | 半导体器件、集成电路各区域或各层的设计制造。 |
| C-Topic-5 | 半导体芯片(包括衬底)中焊盘的设计和互连。 |

### 3.6　结果验证

本文研究数据选取2014－2018年芯片领域专利数据,对挖掘出的技术创新主题有效性采用以下两种科学方式评估:

#### 3.6.1　芯片领域权威专家评估

由于领域内的技术主题在一定程度上相互交叉重叠[34],拟定的技术创新主题也存在一定程度的重叠,单独评分并不符合有效性的准则,专家通过对拟定的技术创新主题进行整合,从整体上对主题进行评分。共邀请了5名微电子领域的专家,他们在芯片技术创新领域具有专业知识和丰富经验。具体评分规则为:1到10分代表技术创新主题的质量由低到高,技术创新主题的质量评分综合3个方面考量:技术价值、创新价值和应用价值,每位微电子专家都对5个拟定的技术创新主题给出自己专业的评判,最后评分如图6所示,评分统计结果见表14。

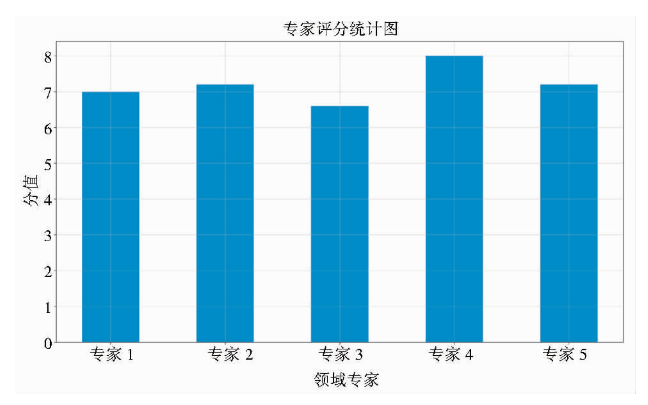



图6　专家评分

#### 3.6.2　最新国内外芯片技术研究

芯片技术跨入后摩尔定律时代[35－36],基于芯片制造环节的深度摩尔定律和基于芯片封装环节的超越摩尔定律已经到来。其中，深度摩尔定律多应用于沟道材料、器件结构、连接导线、高介质金属栅等方面的创新研发,在数字电路中应用较多;超越摩尔定律的主要应用在于将不同模块使用封装技术异质集成在同一封装中,在模拟电路中应用较多。本文挖掘的技术创新主题符合最新芯片技术范畴,其中C-Topic-1、C-Topic-2、C-Topic-3属于芯片封装环节的创新研发主题;C-Topic-4、C-Topic-5属于芯片制造环节的创新研发主题,如表15所示:

表14　专家评分统计

| 序号 | 统计指标 | 分值 |
|---|---|---|
| 1 | 最低分 | 6.6 |
| 2 | 最高分 | 8 |
| 3 | 平均分 | 7.2 |
| 4 | 标准差 | 0.456 |

表15　技术创新主题的芯片技术范畴

| 技术创新主题 | 标记结果 | 芯片技术范畴 |
|---|---|---|
| C-Topic-1 | 围绕半导体衬底的封装技术,包括管芯相对于衬底的位置以及衬底本身的特殊构造,如凹槽等 | 超越摩尔定律 |
| C-Topic-2 | 接触垫在芯片封装中的应用 | 超越摩尔定律 |
| C-Topic-3 | 集成电路或系统中模块的功能配置及互连 | 超越摩尔定律 |
| C-Topic-4 | 半导体器件、集成电路各区域或各层的设计制造 | 深度摩尔定律 |
| C-Topic-5 | 半导体芯片(包括衬底)中焊盘的设计和互连 | 深度摩尔定律 |

## 4　结语

本文提出一种融合多属性的量化方法,快速且有效地挖掘出多个技术创新主题。方法的整体思路是利用LDA挖掘技术主题,然后融合多属性挖掘技术创新主题。表14表明,专家们对挖掘出的技术创新主题的价值是肯定的;表15表明,5个技术创新主题都属于后摩尔时代[35]芯片技术创新研发的方向。根据两种评估方法的结果,可以确定本文提出的融合多属性快速挖掘领域技术创新主题的方法是有效的。本文可能存在的争议是在3.4(2)部分中选取技术创新主题的数量多少上,本文选取5个技术主题作为技术创新主题符合本文的研究目的(快速且有效地挖掘领域内多个技术创新主题),至于"多个技术创新主题"的最优数量,接下来将会通过进一步的研究进行确定。

**参考文献:**

[1] 温军,张森. 专利、技术创新与经济增长[J]. 华东经济管理,2019,33(8):152－158.

[2] SCHMOOKLER J. Changes in industry and in the state of knowl-

edge as determinants of industrial invention[C]// NELSON R R. The rate and direction of inventive activity. Princeton:Princeton University Press, 1962:195 - 232.
[3] GRILICHES Z. Patent statistics as economic indicators: a survey [J]. Journal of economic literature, 1990, 28(4): 1661 - 1707.
[4] SCHMOCH U. Indicators and the relations between science and technology[J]. Scientometrics, 1997, 38(1): 103 - 116.
[5] OECD. Patent statistics manual[M]. Paris: OECD Publishing, 2009.
[6] 赵阳, 文庭孝. 专利技术信息挖掘研究进展[J]. 图书馆, 2018(4): 28 - 33.
[7] CHOI C, PARK Y. Monitoring the organic structure of technology based on the patent development paths[J]. Technological forecasting and social change, 2009, 76(6):754 - 768.
[8] KWON O, SEO J, NOH K, et al. Categorizing influential patents using bibliometric analysis of patent citations network[J]. Information-an international interdisciplinary journal, 2007, 10(3): 313 - 326.
[9] 张欣, 马瑞敏. 基于改进 PageRank 算法的核心专利发现研究[J]. 图书情报工作, 2018, 62(10): 106 - 115.
[10] WANG Y, BAI H J, STANTON M, et al. PLDA: parallel latent dirichlet allocation for large-scale applications[C] //International conference on algorithmic aspects in information and management. San Francisco: Springer-verlag, 2009:301 - 314.
[11] NEWMAN M E J, GIRVAN M. Finding and evaluating community structure in networks[J]. Physical review, 2004, 69(2): 108 - 113.
[12] BLONDEL V D, GUILLAUME J L, LAMBIOTTE R, et al. Fast unfolding of communities in large networks[J]. Journal of statistical mechanics: theory and experiment, 2008, 30(2): 155 - 168.
[13] HAYOUNG C, SEUNGHYUN O, SUNGCHUL C, et al. Innovation topic analysis of technology: the case of augmented reality patents[J]. IEEE access, 2018(6): 16119 - 16137.
[14] 伊惠芳, 吴红, 马永新, 等. 基于 LDA 和战略坐标的专利技术主题分析——以石墨烯领域为例[J]. 情报杂志, 2018, 37(5): 97 - 102.
[15] 范宇, 符红光, 文奕. 基于 LDA 模型的专利信息聚类技术[J]. 计算机应用,2013, 33(S1): 87 - 89, 93.
[16] 吕晓蓉. 专利价值评估指标体系与专利技术质量评价实证研究[J]. 科技进步与对策, 2014, 31(20): 113 - 115.
[17] LANJOUW J, SHANKERMAN M. Stylized facts of patent litigation: value, scopeand ownership[R/OL]. [2019 - 10 - 19]. https://www.nber.org/papers/w6297.pdf.
[18] 孙伟, 刘文静, 葛丽阁, 等. 一种基于词加权 LDA 模型的专利文献分类方法[J]. 计算机技术与发展, 2019(3): 23 - 29.
[19] 庞剑锋, 卜东波, 白硕. 基于向量空间模型的文本自动分类系统的研究与实现[J]. 计算机应用研究, 2001(9): 23 - 26.
[20] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation [J]. Journal of machine learning research, 2003(3): 993 - 1022.
[21] 张晗, 徐硕, 乔晓东. 融合科技文献内外部特征的主题模型发展综述[J]. 情报学报, 2014, 33(10): 1108 - 1120.
[22] 廖列法, 勒孚刚, 朱亚兰. LDA 模型在专利文本分类中的应用[J]. 现代情报, 2017, 37(3): 35 - 39.
[23] 杨超, 朱东华, 汪雪锋. 专利技术主题分析——基于 SAO 结构的 LDA 主题模型方法[J]. 图书情报工作, 2017, 61(3): 86 - 96.
[24] WON S L, EUN J H, SO Y S. Predicting the pattern of technology convergence using big-data on large-scale triadic patents [J]. Technological forecasting & social change, 2015, 100: 317 - 329.
[25] 张文君, 顾行发, 陈良富, 等. 基于均值 - 标准差的 K 均值初始聚类中心选取算法[J]. 遥感学报, 2006, 10(5):715 - 721.
[26] PAOLA D R, SABRINA S, VINCENZO L. A semantic-grained perspective of latent knowledge modeling[J]. Information fusion, 2017, 36: 52 - 67.
[27] 李清海, 刘洋, 吴泗宗, 等. 专利价值评价指标概述及层次分析[J]. 科学学研究, 2007, 25(2): 281 - 286.
[28] 严明义. 函数性数据的统计分析: 思想、方法和应用[J]. 统计研究, 2007(2): 87 - 94.
[29] 温颖, 周昕, 赵文明. 高职软件专业学生职业素养量化评价[J]. 计算机工程与设计, 2017, 38(9): 2586 - 2590.
[30] BIRD S, KLEIN E, LOPER E. Natural language processing with python[M]. New York: O'Reilly Media Press, 2009:41 - 134.
[31] PEDREGOSA F, VAROQUAUX G. Scikit-learn: machine learning in python[J]. Journal of machine learning research, 2011, 12: 2825 - 2830.
[32] LDA Developers. LDA: topic modeling with latent dirichlet allocation [EB/OL]. [2019 - 11 - 22]. https://lda.readthedocs.io/en/latest/.
[33] 国家知识产权局 - 国际专利分类表(2008.01 版)[S/OL]. [2019 - 09 - 01]. http://www.sipo.gov.cn/wxfw/zlwxxxggfw/zsyd/bzyfl/gjzlfl/201406/t20140630_973352.html.
[34] FRANK H, JOCHEN G, MICHAEL H. Memetic search for overlapping topics based on a local evaluation of link communities[J]. Scientometrics, 2016, 11(2): 1089 - 1118.
[35] 张百尚, 商惠敏. 国内外芯片产业技术现状与趋势分析[J]. 科技管理研究, 2019(17): 131 - 134.
[36] 王立娜, 唐川, 房俊民, 等. 2018 年全球半导体领域规划与发展态势分析[J]. 世界科技研究与发展, 2019, 41(2): 120 - 126.

**作者贡献说明:**

李慧: 提出研究思路, 修订论文;

玄洪升: 采集数据, 设计并实现方法, 分析实验结果, 起草论文。

李慧，玄洪升.专利视角下融合多属性的技术创新主题挖掘方法——以芯片领域专利为例[J].图书情报工作,2020,64(11):96－107.

## Multi-attribute Mining Method for Technology Innovation Subject from the Perspective of Patent ——The Case of Chip Patents

Li Hui Xuan Hongsheng
School of Economics and Management, Xidian University, Xi'an 710126

Abstract: [Purpose/significance] By combining multiple attributes, it can quickly and effectively dig out multiple technological innovation themes in the field, providing reference for the determination of technological innovation direction. [Method/process] This paper combined the LDA (Latent Dirichlet Allocation) topic model with the evaluation indicators of patent value, and proposed a quantitative method for mining patent innovation themes. First, TF-IDF, means of perplexity and quartile method were used to construct the LDA topic model of the domain patent to mine technological topics. Then, the probability distribution matrix output by LDA was combined with the evaluation indicators of patent value(claim and IPC) to construct a quantitative indicator system. Then, patents in the chip field were selected for verification experiments, quantitative indicators were calculated and visualized by heat map to identify the technological innovation themes in the field. Finally, based on the mapping relationship between patent, LDA output matrix, innovation theme and quantitative indicators, innovation patent screening and reasonable marking of technological innovation themes were carried out. [Result/conclusion] By inviting experts in the field of microelectronics and based on the latest chip technology at home and abroad to evaluate the experimental results. The scoring results show that the method of mining technology innovation topics with multiple attributes can mine multiple technology innovation topics quickly and effectively. At the practical level, it can better provide ideas for enterprises and scientists in related fields to technological innovation themes.

Keywords: patent perplexity LDA quantitative indicators technological innovation topics